UAV Autonomous Navigation for Wireless Powered Data Collection with Onboard Deep Q-Network

LI Yuting[1], DING Yi[2], GAO Jiangchuan[1], LIU Yusha[1], HU Jie[1], YANG Kun[3]

(1. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. China Mobile Communications Group Jilin Co., Ltd., Changchun 130061, China;
3. School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom)

DOI: 10.12142/ZTECOM.202302011

https://kns.cnki.net/kcms/detail/34.1294.TN.20230518.1444.002.html, published online May 19, 2023

Manuscript received: 2023-02-11

Abstract: In a rechargeable wireless sensor network, utilizing the unmanned aerial vehicle (UAV) as a mobile base station (BS) to charge sensors and collect data effectively prolongs the network's lifetime. In this paper, we jointly optimize the UAV's flight trajectory and the sensor selection and operation modes to maximize the average data traffic of all sensors within a wireless sensor network (WSN) during finite UAV's flight time, while ensuring the energy required for each sensor by wireless power transfer (WPT). We consider a practical scenario, where the UAV has no prior knowledge of sensor locations. The UAV performs autonomous navigation based on the status information obtained within the coverage area, which is modeled as a Markov decision process (MDP). The deep Q-network (DQN) is employed to execute the navigation based on the UAV position, the battery level state, channel conditions and current data traffic of sensors within the UAV's coverage area. Our simulation results demonstrate that the DQN algorithm significantly improves the network performance in terms of the average data traffic and trajectory design.

Keywords: unmanned aerial vehicle; wireless power transfer; deep Q-network; autonomous navigation

Citation (Format 1): LI Y T, DING Y, GAO J C, et al. UAV autonomous navigation for wireless powered data collection with onboard deep Q-network [J]. ZTE Communication, 2023, 21(2): 80 - 87. DOI: 10.12142/ZTECOM.202302011

Citation (Format 2): Y. T. Li, Y. Ding, J. C. Gao, et al., "UAV autonomous navigation for wireless powered data collection with onboard deep Q-Network," *ZTE Communications*, vol. 21, no. 2, pp. 80 - 87, Jun. 2023. doi: 10.12142/ZTECOM.202302011.

1 Introduction

Wireless sensor networks (WSNs) have been widely used in various scenarios, like environment monitoring[1]. However, the energy of the sensors in WSN is usually limited and recharging sensors is challenging[2]. When a WSN is deployed in remote areas, it is not realistic for traditional terrestrial communication networks to charge sensors. In this situation, unmanned aerial vehicles (UAVs) can be used to charge ground sensors and complete tasks such as traffic monitoring, autonomous driving complement, flying relay and data collection[3 - 6]. In addition, various types of natural disasters, such as earthquakes, wildfires, hurricanes, etc., have caused serious damage to communication infrastructure. UAVs as mobile stations can help quickly establish emergency communication and maintain real-time communication to obtain post-disaster situational awareness, which can significantly improve the efficiency of rescue missions.

The UAV is used as a mobile access point (AP) to charge a set of sensors via wireless power transfer (WPT) in the downlink, and the ground sensors leverage the harvested energy to transmit data back to the UAV via wireless information transfer (WIT) in the uplink[7]. However, the resource allocation problem in this scenario is non-convex and difficult to solve directly. Therefore, we formulate the problem as a Markov decision process (MDP), which will be optimally solved with a deep reinforcement learning (DRL) approach[8].

The deep Q-network (DQN) framework has been widely applied in UAV-assisted wireless communication systems. In Ref. [9], the authors investigated UAV-assisted WPT and data collection and employed the DQN to optimize the UAV's instantaneous patrolling velocity as well as plan the flight trajectory, in order to minimize the packet loss. TANG et al.[10] designed a DRL strategy for maximizing the minimum throughput, where the sparse reward was used to ensure that the UAV could complete the optimization task. In Ref. [11], the authors

proposed to dynamically adjust the flying trajectory of the UAV based on the changes of point of interest (PoI) in the coverage range of the UAV, in order to cover as many PoIs as possible, and to improve the fairness of ground users. In Ref. [12], the authors investigated UAV-aided mobile networks, where multiple ground mobile users (GMUs) desired to upload data to a UAV, and maximized the uplink throughput by optimizing the UAV's trajectory. ABEDIN et al.[13] designed a navigation policy for multi-UAVs to improve the data freshness and connectivity to the Internet of Things (IoT) devices, which incorporated different contextual information such as energy and age of information (AoI) constraints. In Ref. [14], the authors investigated a UAV-based emergency communication network, in which UAVs could collect information from ground users in post-disaster scenarios, and transformed the problem into a constrained Markov decision-making process (CMDP). LI et al.[15] formulated a joint optimization of flight cruise control and data collection schedule to minimize network data loss as a partially observable Markov decision process (POMDP), where the states of individual IoT nodes could be obscure to the UAV.

However, the above-mentioned works are all based on the condition that the UAV has partial or full prior knowledge of the environment or waypoints. Investigation into UAVs' navigation with no prior knowledge of sensors' location is still a blank in the literature. Therefore, motivated to fill this gap, we study the UAV navigation problem under the assumption of no prior knowledge about sensor positions on the UAV side. To complete the autonomous navigation task, which is difficult to solve by convex optimization, we propose a novel DRL-based framework to optimize the UAV's trajectory as well as the ground sensors selection with the objective of maximizing the average data traffic of all sensors in a UAV-assisted WSN. The problem is formulated as MDP with a large state and action space. To obtain the up-to-date knowledge about the state information of ground sensors, DRL is used for the UAV to autonomously navigate to the next position. Numerical results show that the proposed algorithm significantly improves the network performance while ensuring the UAV trajectory is optimized.

The rest of this paper is organized as follows. In Section 2, we describe the system model and problem formulation. The DRL-aided algorithm for UAV-assisted networks is presented in Section 3. Section 4 shows the simulation results of the proposed algorithms. Finally, the conclusion of this paper is presented in Section 5.

2 System Model and Problem Formulation

2.1 System Model

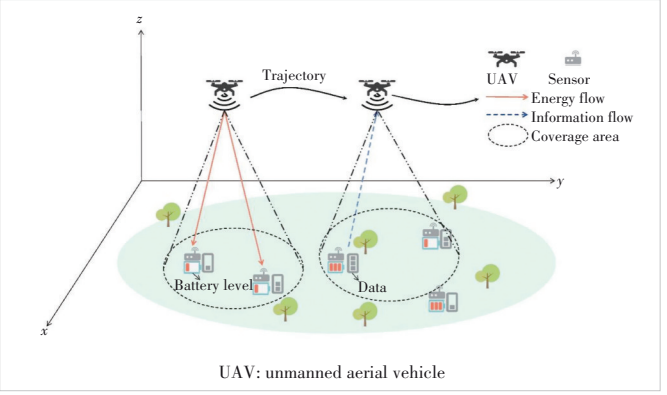
We consider a single-UAV-assisted WSN consisting of K sensors shown in Fig. 1, where the UAV is responsible for charging and collecting data, and the ground sensors harvest energy from the UAV in the downlink and then send collected data in the uplink. Without loss of generality, we assume that the flying height of the UAV is fixed at H m. The UAV has a coverage range of R m^2, i.e., only sensors with a distance of less than R from the UAV can communicate with the UAV. The UAV starts from the origin and flies back to it within the specified flight time. The notations used in this paper are summarized in Table 1.

In actual scenarios, the UAV has no prior knowledge of the location and status of sensors. Therefore, how the UAV performs autonomous navigation also becomes a problem to be solved in our paper. Given the flight height, the UAV coverage area R can be expressed as:

$$R = \pi\left(h\tan\frac{\phi}{2}\right)^2, \tag{1}$$

where ϕ denotes the antenna beam width, and only the sensors in the coverage area can communicate with the UAV.

At any time slot, the UAV has three operation modes: the uplink data collection (DC) mode, the downlink WPT mode and the state listening mode. We employ $\rho_k(t) \in (0,1)$, $\forall k \in \mathcal{K}$ to denote the UAV operational mode selection at the



▲Figure 1. A UAV-assisted rechargeable wireless sensor network (WSN)

▼Table 1. Notation list

Notation	Definition
k, K	Sensor index, number of ground sensors
t, T	Time slot index, total UAV flight time slots
τ, T	Time slot duration, total UAV flight time
$\boldsymbol{w}_k$	Coordinate of the k-th sensor
$v_{\mathrm{UAV}}(t)$	Velocity of the UAV at the t-th time slot
$\boldsymbol{q}(t)$	Position of the UAV at the t-th time slot
$d_k(t)$	Distance between sensor k and the UAV
$h_k(t)$	Channel gain between sensor k and the UAV
R	The coverage range of the UAV
$\bar{P}_k(t)$	Battery level of the k-th sensor
$\rho_k(t)$	Operation mode factor of sensor k
P_{UAV}, P_s	Transmit power of the UAV and ground sensors

UAV: unmanned aerial vehicle

time slot t. In the uplink DC mode, we have $\rho_k(t) = 1$, and sensor k is selected to send its data to the UAV by consuming its energy storage. In order to avoid mutual interference, only a single sensor is allowed to send data to the UAV, which yields $\sum_{k=0}^{K}\rho_k(t) = 1$. In contrast, in the downlink WPT mode, we have $\rho_k(t) = 0$, and the UAV will charge the ground sensors within the coverage area R. These sensors may harvest energy from the downlink radio frequency (RF) signals of the UAV to replenish their energy storage. In the state listening mode, the UAV obtains the status information of the sensors within its coverage area through its beacons, thus making a partial observation of the UAV. The state information includes the sensor's battery level, data traffic, and instantaneous channel conditions. This state information is then used to execute the actions of the UAV. Note that, since the state listening mode occupies a much shorter time duration compared with the other two operation modes, it can be reasonably omitted from the mode selection of the UAV.

Note that the UAV can only obtain state information for sensors within their coverage area R, therefore, the UAV must autonomously navigate to the vicinity of all sensors without fully knowing their locations and should cover as many sensors as possible based on local observations.

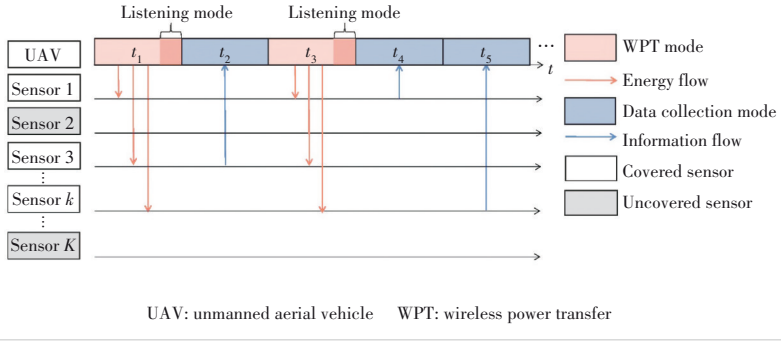
Fig. 2 illustrates the operation modes of the UAV over five consecutive time slots. In the time slots t_1 and t_3, the UAV operates in a downlink WPT mode. Since only sensors 1, 3, and k are in their coverage area, they may harvest energy from the UAV's downlink WPT signal. In the t_2, t_4 and t_5, the UAV operates in an uplink DC mode, while sensors 3, 1 and k upload data to the UAV, respectively.

2.2 Problem Formulation

The locations of sensor k and the UAV at time slot t are denoted as $\boldsymbol{w}_k = (x_k, y_k)$ and $\boldsymbol{q}(t) = (x(t), y(t))$, respectively. Accordingly, the distance from the UAV to sensor k is given by

$$d_k(t) = \sqrt{\left\|\boldsymbol{q}(t) - \boldsymbol{w}_k\right\|^2 + H^2}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector.



▲Figure 2. Communication protocol for a UAV-assisted wireless sensor network

The UAV communicates with sensors via the line-of-sight (LoS) communication links. The channel power gain between the UAV and sensor k at time slot t is given by:

$$h_k(t) = \beta_0 d_k^{-2}(t) = \frac{\beta_0}{\left\|\boldsymbol{q}(t) - \boldsymbol{w}_k\right\|^2 + H^2}, \forall k \in \mathcal{K}, \tag{3}$$

where β_0 denotes the channel power gain at a reference distance of 1 m.

First, we consider the WPT mode, where $\rho_k(t) = 0$. Let P_{UAV} denote the transmit power of the UAV, while all the K sensors have the same transmit power of P_s. Accordingly, at time slot t, the energy harvested by sensor k can be expressed as:

$$\hat{P}_k(t) = \left(1 - \rho_k(t)\right)\eta P_{\mathrm{UAV}} h_k(t)\tau = \frac{\left(1 - \rho_k(t)\right)\eta\beta_0 P_{\mathrm{UAV}}}{\left\|\boldsymbol{q}(t) - \boldsymbol{w}_k\right\|^2 + H^2}\tau, \tag{4}$$

where $0 < \eta \leqslant 1$ denotes the RF-to-direct current energy conversion efficiency.

Then, in the DC mode, we have $\rho_k(t) = 1$. The achievable uplink throughput of sensor k in bits per second can be expressed as:

$$\hat{r}_k(t) = \rho_k(t) B \log_2\left(1 + \frac{P_s h_k(t)}{\sigma^2}\right) = \rho_k(t) B \log_2\left(1 + \frac{P_s h_k(t)}{\left\|q(t) - w_k\right\| + h^2}\right), \tag{5}$$

where σ^2 is the noise power, and $\lambda \triangleq \beta_0/\sigma^2$ is the reference signal-to-noise ratio (SNR). Thus, the total number of data $r_k(t)$ collected from sensor k at the end of the time slot t is given by:

$$r_k(t) = r_k(t-1) + \hat{r}_k(t)\tau, \tag{6}$$

where we reasonably assume $r_k(0) = 0$. By jointly considering energy harvesting and energy consumption at the time slot t, the remaining energy on sensor k is given by

$$P_k(t) = P_k(t-1) + \hat{P}_k(t) - P_s\tau \geqslant 0, \text{for } \forall k. \tag{7}$$

Appropriate actions $\{\alpha(t), v_{\mathrm{UAV}}(t), \rho_k(t)\}$ must be carefully chosen for the UAV to ensure that the energy consumption of sensor k should not exceed the energy stored, which constitutes the energy causality constraint on all sensors of Eq. (7).

For simplicity, in the state listening mode, sensor k may report its energy state to the UAV in the

form of a single binary bit. Therefore, the energy state information of sensor k is quantified as:

$$\bar{P}_k(t)=\begin{cases}1, & \text{if } P_k(t)\geqslant P_s\tau\\ 0, & \text{if } P_k(t)<P_s\tau.\end{cases} \tag{8}$$

Since the flight time is stipulated, we need to predict whether the UAV can fly back to the origin within the specified time at the current position. Accordingly, we express the judgment basis as:

$$\frac{\sum_{t'=1}^{t}v_{\mathrm{UAV}}(t')}{t}(T-t)\geqslant\sqrt{\left\|\boldsymbol{q}(t+1)-\boldsymbol{q}(0)\right\|^2}, \tag{9}$$

where $\frac{\sum_{t'=1}^{t}v_{\mathrm{UAV}}(t')}{t}$ represents the UAV's average flying velocity so far. Additionally, $\boldsymbol{q}(t+1)$ represents the next position obtained by the UAV according to the DQN algorithm that will be introduced in Section 3.

Our objective is to maximize the average data traffic of all sensors during a finite flight time. Therefore, the data traffic optimization problem can be mathematically formulated as follows:

$$(\mathrm{P1}):\max_{\{q(t),(v_{\mathrm{UAV}}),\rho_k(t)\}}\frac{\sum_{k=0}^{K}r_k(T)}{K}, \tag{10}$$

$$\text{s. t. } \rho_k(t)=\{0,1\},\forall k\in\mathcal{K}, \tag{10a}$$

$$\sum_{k=0}^{K}\rho_k(t)=1,\forall t\in T, \tag{10b}$$

$$P_k(t)-P_sT_{\mathrm{DC}}\geqslant 0,\forall k\in\mathcal{K}, \tag{10c}$$

$$r_k(T)\geqslant r_{\mathrm{QoS}},\forall k\in\mathcal{K}, \tag{10d}$$

Eqs. (7) and (9).

In addition to the flight Constraints (9) and (10c), the operating mode Constraints (10a) and (10b) and the energy causality Constraint (7), we also need to guarantee that the average data collection of all sensors should be above the minimum quality of service (QoS) requirement r_{QoS}, as expressed in Constraint (10d).

3 Deep Reinforcement Learning for UAV-Assisted Power Transfer and Data Collection

Due to the non-convex optimization problem and the large action space, in this section, we employ a DQN-based algorithm to solve Problem (P1). The UAV will periodically select the best action based on the network status while maximizing the average data traffic.

3.1 Deep Q-Network

We first provide a brief review of the DQN framework. The DQN approach can be described as a MDP, which is defined by a 4-tuple $\langle\mathcal{S},\mathcal{A},\mathcal{R},\mathcal{P}\rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of all possible actions, $\mathcal{R}$ represents the reward when an action is taken, and $\mathcal{P}$ denotes the transition probability from one state to another. The DQN structure is illustrated in Fig. 3, where the agent observes the environment, obtains the state $s_t\in\mathcal{S}$, chooses an action $a_t\in\mathcal{A}$, and then receives a reward $r_t\in\mathcal{R}$ according the observation and the next state.

The DQN obtains an optimal policy π by maximizing the long-term expected accumulated rewards. The expected accumulated reward for each state-action pair is defined as:

$$Q(s_t,a_t)=\mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\middle|s_t,a_t\right], \tag{11}$$

where $\gamma\in[0,1]$ is the discounted factor. Then, we can get the optimal policy: $\pi(s_t)=\arg\max_{a_t}Q(s_t,a_t)$. By selecting the optimal action a^*, we have the optimal action-value function:

$$\begin{aligned}Q^*(s_t,a_t^*)=&(1-\varpi)Q^*(s_t,a_t^*)+\\&\varpi\left[r(s_t,a_t^*)+\gamma\max_{a_t\in\mathcal{A}}Q(s_t,a_t)\right],\end{aligned} \tag{12}$$

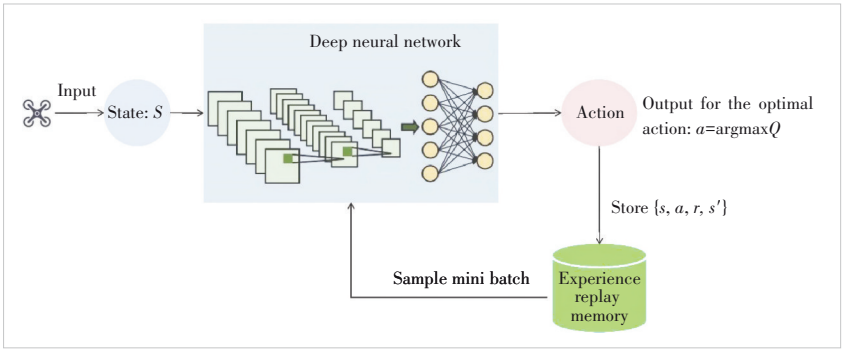
where $\varpi\in(0,1]$ is the learning rate. A DQN uses a deep neural network (DNN) as the approximator of the $Q(\cdot)$ function and is trained by minimizing the following loss function:

$$L(\theta)=\mathbb{E}[y_t-Q(s_t,a_t|\theta)^2], \tag{13}$$

where $\boldsymbol{\theta}^Q$ is the weight vector of the DQN, and y_t is the target value, which is denoted by:

$$y_t=r_t+\gamma\max_{a_t}\hat{Q}(s',a';\theta^-). \tag{14}$$

In conventional Q-learning, the Q-value is updated by both the return at the current slot and the value estimated at the



▲Figure 3. Schematics of the proposed algorithm

next slot. Due to the instability of the training samples, some fluctuations may occur in each iteration, which will be immediately reflected in the next iteration. In order to reduce the impact of related issues, it is necessary to decouple the two parts as much as possible. Therefore, the DQN introduces a dual neural network and replay buffer mechanism. It uses another network to generate the target Q-value, which is used to calculate the evaluated network and loss function. After M iterations, the parameters of the evaluated network will be copied to the target network. The model for calculating the target value will be fixed for a period of time, hence reducing the volatility of the model. In addition, a replay buffer is applied to store the transition samples (s,a,r,s') that are generated in each iteration.

3.2 DQN-Based Solution

In the DQN algorithm we used, the UAV implements the ground sensor selection and operation mode selection, and calculates the next location and flying speed. Each action depends on the network state. Let us explain the definition of state space, action space and reward function of the UAV in our proposed DQN-based algorithm, as summarized in Algorithm 1.

Algorithm 1: DQN-based algorithm

1: Initialize replay memory $\mathcal{D}$ to capacity N;
2: Initialize action-value function Q with random weights θ;
3: Initialize action-value function $\hat{Q}$ with random weights $\theta^{-} = \theta$;
4: **for** episode = 1, M, **do**
5: Start state $s_t \rightarrow s_1$ and accordingly update θ;
6: **for** $t = 1,\cdots,T$ **do**
7: Execute action a_t in an emulator and observe reward r_t and the next state s_{t+1};
8: Store transition (s_t,a_t,r_t,s_{t+1}) in $\mathcal{D}$;
9: Sample M random minibatch of transitions (s_j,a_j,r_j,s_{j+1}), $j = 1,\cdots,M$ from $\mathcal{D}$;
10: Calculate y_j according to Eq. (14);
11: Calculate the MSE loss function $\frac{1}{M}\sum_{j=1}^{M}\left(y_j - Q(s_j,a_j;\theta)\right)^2$ and update the network parameters θ;
12: Update $(x(t+1),y(t+1))$;
13: Every δ steps reset $\hat{Q} = Q$;
14: **end for**
15: **end for**

1) The UAV obtains the status information of the sensors in the coverage area through autonomous navigation. The state $\mathcal{S}$ of the network at the time slot t contains four parts:

• $\bar{P}_k(t)\,|\,k \in \mathcal{K}(t)$: the battery level of the sensors within the coverage area;

• $r_k(t)\,|\,k \in \mathcal{K}(t)$: the current data traffic of all sensors, where $\mathcal{K}(t) = \{k\,|\,d_k(t) \leqslant R\}$, indicates the sensors within the coverage area R of UAV;

• $q(t)$: the UAV location;

• $h_k(t)\,|\,k \in \mathcal{K}(t)$: the channel gain between the UAV and the sensors within the coverage area.

2) The steering angle α_t of the UAV is assumed to be selected from D directions in the angular domain. The action $\mathcal{A}$ contains four parts:

• k: the ground sensor selection, where $k \in \mathcal{K}(t)$;

• $\rho_k(t)$: the operation mode selection, by which the UAV chooses to charge or receive data, where $\rho_k(t) \in \{0,1\}$;

• $\alpha(t)$: the steering angle;

• $v_{\mathrm{UAV}}(t)$: the next flying speed.

As it can be seen from the above, the action space $\mathcal{A}$ has a cardinality of $3 \cdot K \cdot D$, where 3 represents that the flying speed of the UAV can have three values. Thus we can get the next location of the UAV:

$$\begin{cases} x(t+1) = x(t) + v_{\mathrm{UAV}}\cos\alpha(t), \\ y(t+1) = y(t) + v_{\mathrm{UAV}}\sin\alpha(t). \end{cases} \tag{15}$$

3) The reward $\mathcal{R}$: since our objective is to maximize the average data traffic while ensuring that the battery level $\bar{P}_k(t)$ is not lower than 1, our reward function consists of two parts:

• $r_{\mathrm{data}}(t)$ is the change of the average data traffic of all sensors after selecting an action:

$$r_{\mathrm{data}}(t) = \Delta\frac{\sum_{k=1}^{K} r_k(t)}{K}. \tag{16}$$

• $r_{\mathrm{penalty}}(t)$ is an action penalty when none of the constraints in (P1) is satisfied.

To summarize, we give the final reward function:

$$r(t) = r_{\mathrm{data}}(t) + r_{\mathrm{penalty}}(t). \tag{17}$$

4 Simulation Results

In this section, we present network configurations and illustrate numerical results including the trajectory of the UAV, energy, battery level and data traffic of all sensors to validate the proposed DQN-based algorithm. Our experiments are performed on Tensorflow 1.11.0 (the symbolic math library for numerical computation) and Python 3.6.

4.1 Experiments Settings

As it is shown in Tables 2 and 3, we assume that K =10 sensors are uniformly distributed within a 200 × 200 m^2 district in our environments. The time-slot duration is fixed at t = 1 s. The transmit power of the UAV and ground sensors are P_{UAV} = 30 dBm and P_s = 0 dBm, respectively. The channel power gain at the reference distance d_0 = 1 m is set as β_0 = −30 dB. The QoS requirement is set as 0 bit/s. And we set

▼Table 2. Simulation parameters: DNN

DNN Parameters	Value
Learning rate	0.000 1
Discount factor	0.9
Replay memory size	10 000
Batch size	32
ReLu hidden neurons	20
Number of neural network layers	2

DNN: deep neural network

▼Table 3. System parameters

System Parameters	Value
Bandwidth	1 MHz
Energy conversion efficiency	0.9
Noise power	−60 dBm
Flying height	10 m
Coverage area	70 m^2
Steering angle $(\alpha(t))$	$\left\{\frac{k}{4}\pi \mid k = 0,\cdots,7\right\}$
Flying velocity $(v_{UAV}(t))$	{0,5,10} m/s

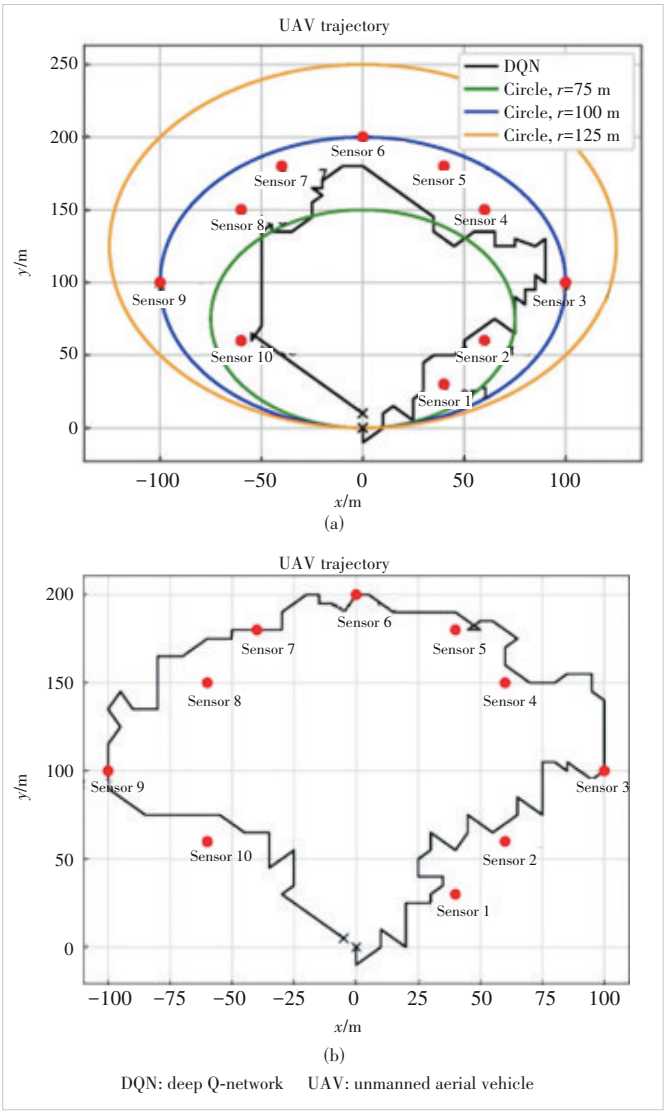
$r_{penalty}$ as 0.01 when a constraint is not satisfied.

4.2 Performance Evaluation

The optimized trajectory of the UAV within different specified flight time is shown in Fig. 4. The red dots represent 10 sensors, and the asterisks indicate the start and end of the UAV. The black trajectories represent the UAV trajectories obtained by DQN. We assume that when the distance between the UAV and the origin is less than 10 m and all sensors satisfy the QoS requirement, the episode is ended. It can be observed that the UAV can complete autonomous navigation tasks and adjust its trajectory either to transfer energy or collect data. As it can be seen from Fig. 4, the UAV needs to cover as many sensors as possible for charging in order to receive data later, so the UAV will not fly straightly over a specific UAV. Similarly, we can see from Figs. 4(a) and 4(b) that when the flight time increases, the UAV's flight trajectory is closer to the sensor.

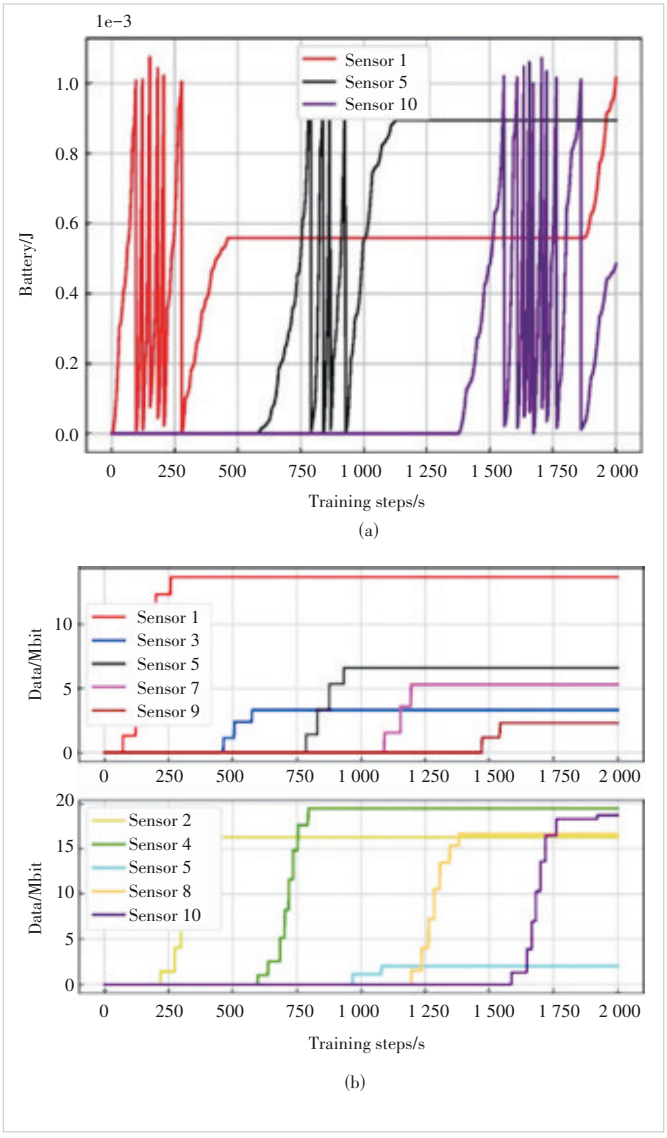
Furthermore, we compare the situation of the UAV flying along the preset circular trajectory with the same start points. As shown in the colored trajectories in Fig. 4(a), the UAV will fly at a constant speed along circular trajectories with the radius of 75 m, 100 m, and 125 m respectively. At this time, the action of the UAV will only include the selection of the operation mode and the specific sensor, namely $a_t = \{k, \rho_k(t)\}$.

We then investigate the DQN-based algorithm performance by analyzing the battery and data traffic of sensors. Figs. 5(a) and 5(b) respectively show the energy and data traffic of the first, fifth and tenth sensors, where a training step is equivalent to the time slot duration. The rising part of the curve in Fig. 5(a) indicates that the corresponding sensor is being charged, which corresponds to the level part of the curve in Fig. 5(b). In contrast, the falling part of Fig. 5(a) indicates that the sensor k transmits data to the UAV, which also corresponds to the rising curve in Fig. 5(b).

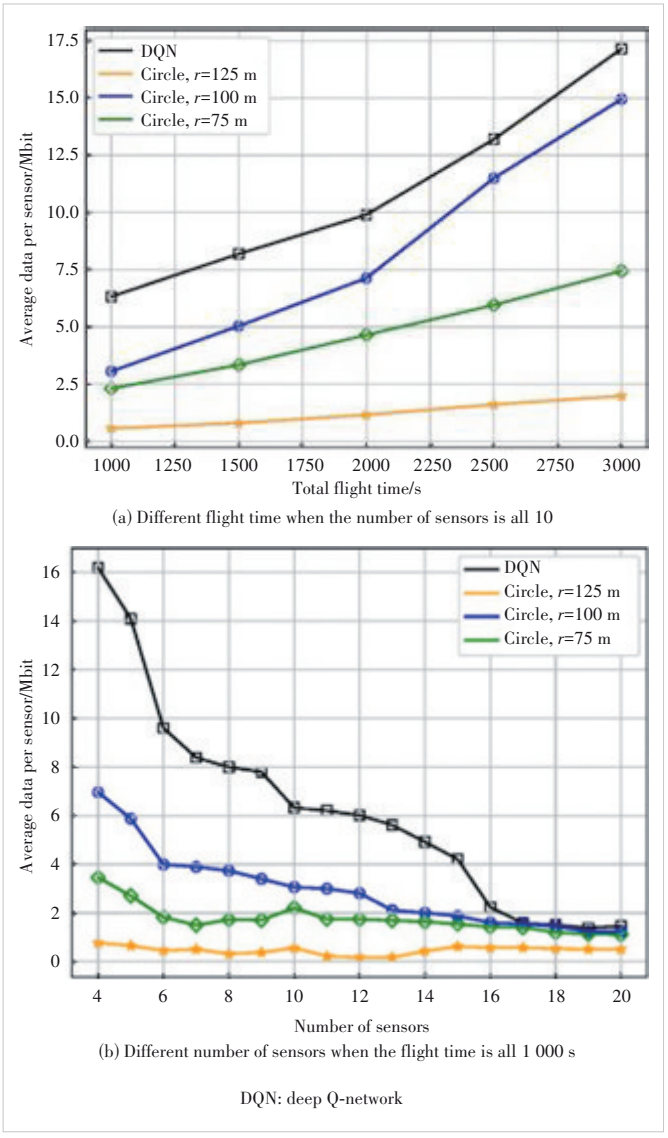


▲ Figure 4. UAV's trajectory under different flight time: (a) 2 000 s and (b) 4 000 s

For the purpose of testing the performance of the algorithm, we then simulate the scene with different sensor numbers and different flight time by DQN and three circular trajectories with different radii in Fig. 6. It can be seen from Fig. 6(a) that when the flight time of UAV increases, the average data traffic on the sensor increases. This is because the flight time increases, and the UAV will have more time to receive data from the sensors. And it can be seen that the performance obtained by the DQN algorithm is the best, followed by the circular trajectory with r = 100 m, and the circular trajectory with r = 125 m is the worst. Because the DQN algorithm can choose the speed and the next position of the UAV, it can hover around the sensor more. The circu-



▲ **Figure 5. Energy and data traffic of k=1.5 and 10 ground sensors: (a) battery and (b) data traffic**



▲ **Figure 6. Comparison of average data traffic per sensor by DQN and three circular trajectories**

lar trajectory with r = 100 m is the closest to the sensor among all circular trajectories shown in Fig. 4(a), so the average data traffic obtained is the largest among the three circular trajectories.

Fig. 6(b) shows that under the same flight time, the fewer the number of sensors, the more the UAV will choose to perform the data collection mode, which results in higher average data traffic. After the number of sensors increases to 16, the average data traffic per sensor by DQN becomes stabilized. This is because the total distribution area of the sensors and the coverage of the UAV remains unchanged. When the number of sensors increases, the distribution of sensors will be denser, and the UAV can charge more sensors with the same coverage and the same charging time. Additionally, it can be observed from Fig. 6(b) that the number of sensors does not have much impact on the circular trajectories, since the UAV always flies at a constant speed. There is no situation where the UAV has a higher probability of hovering over certain sensors when the number of sensors is small. This shows that the DQN algorithm is more suitable for sparsely distributed and nonuniform networks.

5 Conclusions

In this paper, we formulate the problem of power transfer and data collection in a UAV-assisted WSN, while ensuring autonomous navigation of the UAV. We then propose a DQN-based algorithm to solve the problem in order to maximize the average data traffic. Given the states of the battery level, the current data traffic of all sensors, the position of the UAV, and channel conditions, the UAV takes actions according to the

proposed DQN-based algorithm. Numerical results illustrate that the proposed algorithm significantly improves the network performance. However, the proposed algorithm also has high complexity. Due to the real-time interaction with the environment, the UAV needs significant energy consumption, which can be solved by the digital twin network.

References

[1] HU J, WANG Q, YANG K. Energy self-sustainability in full-spectrum 6G [J]. IEEE wireless communications, 2021, 28(1): 104 – 111. DOI: 10.1109/MWC.001.2000156
[2] WU D P, HE J, WANG H G, et al. A hierarchical packet forwarding mechanism for energy harvesting wireless sensor networks [J]. IEEE communications magazine, 2015, 53(8): 92 – 98. DOI: 10.1109/MCOM.2015.7180514
[3] GHARIBI M, BOUTABA R, WASLANDER S L. Internet of drones [J]. IEEE access, 2016, 4: 1148 – 1162. DOI: 10.1109/ACCESS.2016.2537208
[4] ZHU S C, GUI L, CHENG N, et al. Joint design of access point selection and path planning for UAV-assisted cellular networks [J]. IEEE Internet of Things journal, 2020, 7(1): 220 – 233. DOI: 10.1109/JIOT.2019.2947718
[5] BEN GHORBEL M, RODRÍGUEZ-DUARTE D, GHAZZAI H, et al. Joint position and travel path optimization for energy efficient wireless data gathering using unmanned aerial vehicles [J]. IEEE transactions on vehicular technology, 2019, 68(3): 2165 – 2175. DOI: 10.1109/TVT.2019.2893374
[6] HU J, CAI X P, YANG K. Joint trajectory and scheduling design for UAV aided secure backscatter communications [J]. IEEE wireless communications letters, 2020, 9(12): 2168 – 2172. DOI: 10.1109/LWC.2020.3016174
[7] XIE L F, XU J, ZHANG R. Throughput maximization for UAV-enabled wireless powered communication networks [J]. IEEE Internet of Things journal, 2019, 6 (2): 1690 – 1703. DOI: 10.1109/JIOT.2018.2875446
[8] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning [EB/OL]. [2023-03-10]. https://arxiv.org/abs/1312.5602
[9] LI K, NI W, TOVAR E, et al. On-board deep Q-network for UAV-assisted online power transfer and data collection [J]. IEEE transactions on vehicular technology, 2019, 68(12): 12215 – 12226. DOI: 10.1109/TVT.2019.2945037
[10] TANG J, SONG J R, OU J H, et al. Minimum throughput maximization for multi-UAV enabled WPCN: A deep reinforcement learning method [J]. IEEE access, 2020, 8: 9124 – 9132. DOI: 10.1109/ACCESS.2020.2964042
[11] LIU C H, CHEN Z Y, TANG J, et al. Energy-efficient UAV control for effective and fair communication coverage: a deep reinforcement learning approach [J]. IEEE journal on selected areas in communications, 2018, 36(9): 2059 – 2070. DOI: 10.1109/JSAC.2018.2864373
[12] LU Y P, XIONG G, ZHANG X, et al. Uplink throughput maximization in UAV-aided mobile networks: a DQN-based trajectory planning method [J]. Drones, 2022, 6(12): 378. DOI: 10.3390/drones6120378
[13] ABEDIN S F, MUNIR M S, TRAN N H, et al. Data freshness and energy-efficient UAV navigation optimization: a deep reinforcement learning approach [J]. IEEE transactions on intelligent transportation systems, 2021, 22 (9): 5994 – 6006. DOI: 10.1109/TITS.2020.3039617
[14] ZHANG T K, LEI J Y, LIU Y W, et al. Trajectory optimization for UAV emergency communication with limited user equipment energy: a safe-DQN approach [J]. IEEE transactions on green communications and networking, 2021, 5(3): 1236 – 1247. DOI: 10.1109/TGCN.2021.3068333
[15] LI K, NI W, TOVAR E, et al. Joint flight cruise control and data collection in UAV-aided Internet of Things: an onboard deep reinforcement learning approach [J]. IEEE Internet of Things journal, 2021, 8(12): 9787 – 9799. DOI: 10.1109/JIOT.2020.3019186

Biographies

LI Yuting is with the University of Electronic Science and Technology of China. Her current research interests include data and energy integrated communication networks and machine learning.

DING Yi received his bachelor's degree in communication and master's degree in communication and information systems engineering both from Beijing University of Posts and Telecommunications, China. After graduation, he has been engaged in the field of wireless network planning and optimization, demonstrating expertise in the 5G network technology and its evolving trends.

GAO Jiangchuan received his BS degree from the Southwest Jiaotong University, China in 2022. He is currently pursuing his master's degree in the University of Electronic Science and Technology of China. His current research interests include data and energy integrated communication networks and unmanned aerial vehicle trajectory planning.

LIU Yusha (yusha.liu@uestc.edu.cn) received her PhD degree from the University of Southampton, UK, and is currently with the University of Electronic Science and Technology of China. Her current research interests include wireless communications, signal processing and deep learning.

HU Jie received his BE and MS degrees from Beijing University of Posts and Telecommunications, China in 2008 and 2011, respectively, and received his PhD degree from the School of Electronics and Computer Science, University of Southampton, UK in 2015. Since 2016, he has been working with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). He is now a research professor and PhD supervisor. He won UESTC's Academic Young Talent Award in 2019. Now he is supported by the "100 Talents" program of UESTC. He is an editor for *IEEE Wireless Communications Letters*, *IEEE/CIC China Communications* and *IET Smart Cities*. He serves for *IEEE Communications Magazine*, *Frontiers in Communications and Networks* as well as *ZTE communications* as a guest editor. He is a program vice-chair for IEEE TrustCom 2020, a technical program committee (TPC) chair for IEEE UCET 2021 and a program vice-chair for UbiSec 2022. He also serves as a TPC member for several prestigious IEEE conferences. He has won the best paper award of IEEE SustainCom 2020 and the best paper award of IEEE MMTC 2021. His current research focuses on wireless communications and resource management for B5G/6G, wireless information and power transfer as well as integrated communication, computing and sensing.

YANG Kun received his PhD from the Department of Electronic & Electrical Engineering of University College London (UCL), UK. He is a Chair Professor in the School of Computer Science & Electronic Engineering, University of Essex, UK, and is leading the Network Convergence Laboratory (NCL), UK. He is also an affiliated professor at UESTC, China. Before joining in the University of Essex at 2003, he had worked at UCL on several European Union (EU) research projects for several years. His main research interests include wireless networks and communications, IoT networking, data and energy integrated networks and mobile computing. He manages research projects funded by various sources such as UK EPSRC, EU FP7/H2020, etc. He has published more than 400 journal papers and filed 20 patents. He is an IEEE ComSoC Distinguished Lecturer (2020 – 2021) and a member of Academia Europaea (MAE). Professor YANG is an IEEE Fellow.